REPORT

Alignment Analysis of Mathematics Standards and Assessments

Missouri

Grades 3–8 and 10
2006 and 2007

Norman L. Webb

November 29, 2006

REPORT

Alignment Analysis of Mathematics Standards and Assessments

Missouri
Grades 3–8 and 10
2006 and 2007

Norman L. Webb

November 29, 2006

Acknowledgements

Reviewers

Lynn Raith	(Group Leader)	PA
Linda McQuillen		WI
Tom Loomis		WI
Brian Vesperman		WI
Karen Pace		MO
Michelle Hofmann		MO
Pamela M. Herd		MO
Curtis Wren		MO

The State of Missouri funded this analysis. Walt Brown, Coordinator of Curriculum and Assessment, Michael Muenks, Director of Assessment, and Susan Newbold, Assessment Supervisor, were the main contact people for the Missouri Department of Elementary and Secondary Education, Division of School Improvement, which oversaw the coordination of the study.

Table of Contents

Executive Summary

Eight reviewers analyzed the mathematics assessments for 2006 and 2007 and the Missouri Standards at a three-day institute held September 27-29, 2006, in Columbia, Missouri. The reviewers included mathematics content experts, district mathematics supervisors, mathematics teachers, and mathematics education doctoral students. Half of the reviewers were from Missouri and half were from other states. All eight reviewers analyzed the depth-of-knowledge level of the grade-level expectations under each of the five standards for all grades 3–8 and 10. At least four reviewers analyzed the assessment for each year. The intraclass correlation among the reviewers across all 14 analyses was reasonably high, along with the pairwise agreement in assigning items to standards. However, pairwise agreement in the assignment of items to GLEs was lower than for other alignment studies. This lower level of agreement could be the result of lack of clarity in the GLE statements, or the need for more training of the reviewers.

For both years, 2006 and 2007, the alignment was reasonable. The alignment needed slight improvement over only one grade for each year—grade 5 for 2006 and grade 10 for 2007. From five to seven items would need to be replaced or added to achieve full alignment on these two test forms. For all other grades for each of the years, the assessment and the standards were either fully aligned or reasonably aligned. For these assessment forms and grades, less than five items would need to be replaced or added. The assessment and standards were found to be fully aligned for grades 4 and 6 (2006) and for grades 3 and 4 (2007). All test forms for all grades and years had a sufficient number of items for each standard that were, in general, adequately distributed among the GLEs. The main alignment issues were that one or two of the items had too low a depth-of-knowledge level compared to the DOK level of the corresponding GLE. Also, for a few standards, one GLE was overemphasized compared to the other GLEs under that standard. A relatively higher number of items were coded to generic GLEs, indicating a lack of precise match of the item to a specific GLE. Most of these items were on the norm-referenced part of the assessment, Session 2, which would be more difficult to change. Considering all seven grades and both years, the alignment is judged to be reasonable.

Alignment Analysis of Mathematics Standards and Assessments

Missouri
Grades 3–8 and 10

Norman L. Webb

Introduction

The alignment of expectations for student learning with assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. Alignment is defined as the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do. As such, alignment is a quality of the relationship between expectations and assessments and not an attribute of any one of these two system components. Alignment describes the match between expectations and an assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, alignment is determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997).

A three-day Alignment Analysis Institute was conducted September 20-22, 2006, in Columbia, Missouri. Eight reviewers, including mathematics content experts, district mathematics supervisors, mathematics teachers, and mathematics education doctoral students analyzed the agreement between the state's mathematics standards and the 2006 assessments for grades 3–8 and 10. Four of the reviewers were from Missouri and four were from other states.

The State of Missouri uses the terminology of *standards* and *grade-level expectations* in its mathematics content expectations. Standards were the broad content requirements across all grades. Five mathematics standards were included in the analysis. Grade-level expectations (sometimes referred to as objectives) specified in greater detail under a standard what students are to know and do. The grade-level expectations were clustered under an intermediary level of expectations defined for the purpose of this analysis as *goals*. For example, the standard, Number and Operations, was divided into three goals (e.g., Understand numbers, ways of representing numbers, relationships among numbers, and number systems). The goals were specified by two to four grade-level expectations (e.g., Read, write, and compare whole numbers up to 3 digits). Data for this analysis were entered for the grade-level expectations and reported out at the standards level.

As part of the alignment institute, reviewers were trained to identify the depth-of-knowledge of the grade-level expectations and assessment items. This training included

reviewing the definitions of the four depth-of-knowledge (DOK) levels and reviewing examples of each. Then the reviewers participated in 1) a consensus process to determine the depth-of-knowledge levels of the grade-level expectations and 2) individual analyses of the assessment items. Following individual analyses of the items, reviewers participated in a debriefing discussion in which they assessed the degree to which they had coded particular items or types of content to the grade-level expectations.

Assessments for 2006 and 2007 were analyzed in this study. At least four reviewers analyzed each assessment, two Missouri reviewers and two external reviewers. A Latin square scheme was used to rotate reviewers so that different combinations of reviewers analyzed each form. This was done to minimize error that may be associated with any one reviewer.

To derive the results from the analysis, the reviewers' responses are averaged. Any variance among reviewers is considered legitimate, with the true depth-of-knowledge level for the item falling somewhere between two or more assigned values. Such variation could signify a lack of clarity in how the standards and grade-level expectations were written, the robustness of an item that can legitimately correspond to more than one grade-level expectation, and/or a depth of knowledge that falls in between two of the four defined levels. Reviewers were allowed to identify one assessment item as corresponding to up to three grade-level expectations—one primary hit (grade-level expectation) and up to two secondary hits. However, reviewers could only code one depth-of-knowledge level to each assessment item, even if the item corresponded to more than one grade-level expectation.

Reviewers were instructed to focus primarily on the alignment between the state standards and assessments. However, reviewers were encouraged to offer their opinions on the quality of the standards, or of the assessment activities/items, by writing a note about the item. Reviewers could also indicate whether there was a Source-of-Challenge issue with the item—i.e., a problem with the item that might cause the student who knows the material to give a wrong answer, or enable someone who does not have the knowledge being tested to answer the item correctly.

The results produced from the institute pertain only to the issue of alignment between the Missouri state standards and the state assessment instruments. Note that this alignment analysis does not serve as external verification of the general quality of the state's standards or assessments. Rather, only the degree of alignment is discussed in the results. For these results, the means of the reviewers' coding were used to determine whether the alignment criteria were met. When reviewers did vary in their judgments, the means lessened the error that might result from any one reviewer's finding. Standard deviations are reported in the tables provided in the Appendix, which give one indication of the variance among reviewers.

The present report describes the results of an alignment study of grade-level expectations and both the 2006 and 2007 operational tests in mathematics for grades 3–8 and 10 in Missouri. The study addressed specific criteria related to the content agreement

between the state standards and grade-level assessments. Four criteria received major attention: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation.

Alignment Criteria Used for This Analysis

This analysis judged the alignment between the standards and the assessments on the basis of four criteria. Information is also reported on the quality of items by identifying items with Sources-of-Challenge and other issues. For each alignment criterion, an acceptable level was defined by what would be required to assure that a student had met the standards.

Categorical Concurrence

An important aspect of alignment between standards and assessments is whether both address the same content categories. The categorical-concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. *The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents.* This criterion was judged by determining whether the assessment included items measuring content from each standard. The analysis assumed that the assessment had to have at least six items for measuring content from a standard in order for an acceptable level of categorical concurrence to exist between the standard and the assessment. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. Of course, many factors have to be considered in determining what a reasonable number is, including the reliability of the subscale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63. This indicates that about 63% of the group would be consistently classified as masters or nonmasters if two equivalent test administrations were employed. The agreement coefficient would increase if the cutoff score is increased to one standard deviation from the mean to .77 and, with a cutoff score of 1.5 standard deviations from the mean, to .88. Usually states do not report student results by standards or require students to achieve a specified cutoff score on subscales related to a standard. If a state did do this, then the state would seek a higher agreement coefficient than .63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a standard, and as a basis for making some decisions about students' knowledge of that standard. If the mean for six items is 3 and one standard deviation is one item, then a cutoff score set at 4 would produce an agreement coefficient of .77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale.

Depth-of-Knowledge Consistency

Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-knowledge consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.* For consistency to exist between the assessment and the standard, as judged in this analysis, at least 50% of the items corresponding to a standard had to be at or above the level of knowledge of the standard: 50%, a conservative cutoff point, is based on the assumption that a minimal passing score for any one standard of 50% or higher would require the student to successfully answer at least some items at or above the depth-of-knowledge level of the corresponding standard. For example, assume an assessment included six items related to one standard and students were required to answer correctly four of those items to be judged proficient—i.e., 67% of the items. If three, 50%, of the six items were at or above the depth-of-knowledge level of the corresponding standards, then for a student to achieve a proficient score would require the student to answer correctly at least one item at or above the depth-of-knowledge level of one standard. Some leeway was used in this analysis on this criterion. If a standard had between 40% and 50% of items at or above the depth-of-knowledge levels of the standards, then it was reported that the criterion was "weakly" met.

Interpreting and assigning depth-of-knowledge levels to both grade-level expectations within standards and to assessment items is an essential requirement of alignment analysis. These descriptions help to clarify what the different levels represent in mathematics:

Level 1 (Recall) includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify Level 1 include "identify," "recall," "recognize," "use," and "measure." Verbs such as "describe" and "explain" could be classified at different levels, depending on what is to be described and explained.

Level 2 (Skill/Concept) includes the engagement of some mental processing beyond an habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret," could be classified at different levels depending on the object of the action. For example, interpreting information from a

simple graph, or requiring mathematics information from the graph, also is at Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is at Level 3. Level 2 activities are not limited solely to number skills, but can involve visualization skills and probability skills. Other Level 2 activities include noticing and describing non-trivial patterns; explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be at Level 3. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve problems.

Level 4 (Extended Thinking) requires complex reasoning, planning, developing, and thinking, most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be at Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas *within* the content area or *among* content areas—and to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include developing and proving conjectures; designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

Range-of-Knowledge Correspondence

For standards and assessments to be aligned, the breadth of knowledge required on both should be comparable. *The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.* The criterion for correspondence between span of knowledge for a standard and an assessment considers the number of grade-level

expectations within the standard with one related assessment item/activity. Fifty percent of the grade-level expectations for a standard had to have at least one related assessment item in order for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a standard. This assumes that each benchmark for a standard should be given equal weight. Depending on the balance in the distribution of items and the need to have a low number of items related to any one grade-level expectation, the requirement that assessment items need to be related to more than 50% of the grade-level expectations for a standard increases the likelihood that students will have to demonstrate knowledge on more than one grade-level expectation per standard to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring an assessment to include items related to a greater number of the grade-level expectations. However, any restriction on the number of items included on the test will place an upper limit on the number of grade-level expectations that can be assessed. Range-of-knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of standards and a large number of grade-level expectations. If 50% or more of the grade-level expectations for a standard had a corresponding assessment item, then the range-of-knowledge correspondence criterion was met. If between 40% and 50% of the grade-level expectations for a standard had a corresponding assessment item, the criterion was "weakly" met.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned standards and assessments require that knowledge be distributed equally in both. The range-of-knowledge criterion only considers the number of grade-level expectations within a standard hit (a standard with a corresponding item); it does not take into consideration how the hits (or assessment items/activities) are distributed among these grade-level expectations. *The balance-of-representation criterion is used to indicate the degree to which one grade-level expectation is given more emphasis on the assessment than another.* An index is used to judge the distribution of assessment items. This index only considers the grade-level expectations for a standard that have at least one hit—i.e., one related assessment item per grade-level expectation. The index is computed by considering the difference in the proportion of grade-level expectations and the proportion of hits assigned to the grade-level expectation. An index value of 1 signifies perfect balance and is obtained if the hits (corresponding items) related to a standard are equally distributed among the grade-level expectations for the given standard. Index values that approach 0 signify that a large proportion of the hits are on only one or two of all of the grade-level expectations hit. Depending on the number of grade-level expectations and the number of hits, a unimodal distribution (most items related to one grade-level expectation and only one item related to each of the remaining grade-level expectations) has an index value of less than .5. A bimodal distribution has an index value of around .55 or .6. Index values of .7 or higher indicate that items/activities are distributed among all of the grade-level expectations at least to some degree (e.g., every grade-level expectation has at least two items) and is used as the acceptable level on this

criterion. Index values between .6 and .7 indicate the balance-of-representation criterion has only been "weakly" met.

Source-of-Challenge Criterion

The Source-of-Challenge criterion is only used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted mathematics grade-level expectation, concept, or application. Cultural bias or specialized knowledge could be reasons for an item to have a Source-of-Challenge problem. Such item characteristics may result in some students not answering an assessment item, or answering an assessment item incorrectly, or at a lower level, even though they possess the understanding and skills being assessed.

Findings – 2006 Study

Standards

Eight reviewers participated in the depth-of-knowledge (DOK) level consensus process for the standards and grade-level expectations for the Missouri Mathematics Standards. A summary of their deliberations is presented in Table 1. The complete group consensus values for each standard and grade-level expectation (GLE) can be found in Appendix A. The reviewers judged that the grade-level expectations were primarily at the skills and concept level of complexity. The proportion of GLEs with a DOK level of 2 varied between 66% to 86% over the seven grades. In general, the distribution of DOK levels across the grades was relatively flat, with little increase in sophistication at the higher grade levels. Reviewers judged that 18% of the grade 5 grade-level expectations were at DOK Level 3 (Strategic reasoning), the highest percentage for this level for all seven grades.

The reviewers were told that within each standard (e.g., Numbers and Operations), the grade-level expectations were intended to fully span the content of that standard. For this reason, the reviewers only coded items to a standard if there were no grade-level expectation that the item appeared to target. Such items are considered to target a generic grade-level expectation. A large number of items coded to generic grade-level expectations may indicate ways in which a standard's content is not fully spanned or described by its grade-level expectations. This may also simply indicate that these items are not as precise as the grade-level expectations. Table 2 shows the items on each of the seven assessments that were coded to a generic grade-level expectation by more than one reviewer.

Table 1
Percent of Grade-level Expectations by Depth-of-Knowledge (DOK) Levels for Grades 3–8 and 10, Missouri Alignment Analysis for Mathematics 2005 and 2006 Study

Grade	Total Number of Grade-level Expectations (GLEs)	DOK Level	# of objs by Level	% within std by Level
3	32	1 2 3	5 24 3	15 75 9
4	34	1 2 3	4 27 3	11 79 8
5	33	1 2 3	5 22 6	15 66 18
6	34	1 2 3	6 24 4	17 70 11
7	38	1 2 3	8 27 3	21 71 7
8	36	1 2 3	7 24 5	19 66 13
10	36	1 2 3	1 31 4	2 86 11

The reviewers were told that within each standard (e.g., Numbers and Operations), the grade-level expectations were intended to fully span the content of that standard. For this reason, the reviewers only coded items to a standard if there were no grade-level expectation the item appeared to target. Such items are considered to target a generic grade-level expectation. A large number of items coded to generic grade-level expectations may indicate ways in which a standard's content is not fully spanned or described by its grade-level expectations. This may also simply indicate that these items are not as precise as the grade-level expectations. Table 2 shows the items on each of the seven assessments that were coded to a generic grade-level expectation by more than one reviewer.

A relatively large number of items were coded by two or more reviewers to generic grade-level expectations on the 2006 and 2007 mathematics assessments (Table 2). However, most of these items were in Session 2 of the assessments, the norm-referenced part of the assessments. The Session 1 items are shaded in Table 2. For grades

3, 5, 6, and 7, Session 2 began with item number 28. For grades 4 and 8, Session 2 began with item number 31. For grade 10, Session 2 began with item number 33.

Across all grades, the number of items coded by two or more reviewers to a generic GLE ranged from 7 to 18 items. However, the number of items coded to Session 1 items ranged from 0 (grade 6) to 5 (grade 4). Two or more reviewers did not find a good match between the grade-level expectations and the assessment items for 18 grade 5 items, 15 grade 6 items, and 15 grade 10 items. For many of these items, the reviewers indicated in their notes (Appendix C) that the item matched a grade-level expectation in a previous grade. For other items, the content required by the item was not explicitly stated in the grade-level expectations. For example, Item 12 on the grade 10 2006 mathematics form assessed students' knowledge of perimeter and circumference, whereas the grade-level expectation M.2.c.10 (the GLE that most likely fit) only addressed three-dimensional objects.

The Missouri items coded to the generic grade-level expectations should be reviewed to confirm whether a precise match exists between a grade-level expectation and the item. The largest number of Missouri items assigned to generic GLEs were at grades 3 and 4. The alignment issue may be with the wording of the grade-level expectation, rather than with the assessment. The reviewers' notes and source-of-challenge comments frequently will point to what the reviewer identified as the issue. Based on this analysis, from 5% to 20% of the items on the assessment do not precisely match what reviewers judged as the intended grade-level expectation.

Table 2
Items Coded to Generic Grade-level Expectations by More Than One Reviewer, Missouri Alignment Analysis for Mathematics, Grades 3–8 and 10, 2006 and 2007 Assessments

Grade	2006		2007	
	Assessment Item Number	Generic GLE (Number of Reviewers)	Assessment Item Number	Generic GLE (Number of Reviewers)
3	58	A.1 (2)		
3	15	G.1 (2)		
3	21	M.2 (2)		
3	7	D.4 (2)		
3			12	G.1 (2)
3			37	M.1 (2)
3			47	M.1 (2)
4	36	N.1 (2)	36	N.1 (2)
4			33	N. 3 (2)
4			40	N. 3 (3)
4	48	N.3 (2)	48	N. 3 (2)
4	57	N.3 (2)	57	N. 3 (3)
4			30	A.3 (3)

Table 2 (continued)
Items Coded to Generic Grade-level Expectations by More Than One Reviewer, Missouri Alignment Analysis for Mathematics, Grades 3–8 and 10, 2006 and 2007 Assessments

Grade	2006		2007	
	Assessment Item Number	Generic GLE (Number of Reviewers)	Assessment Item Number	Generic GLE (Number of Reviewers)
4			53	M.1 (2)
4			18	M.2 (3)
4			27	M.2 (3)
4			61	M.2 (2)
4			26	D.1 (2)
4			12	D.4 (3)
5	36	N.1 (3)		
5	46	N.1 (2)		
5	28	N.3 (4)	28	N.3 (3)
5	30	N.3 (3)	30	N.3 (4)
5	31	N.3 (4)	31	N.3 (3)
5	33	N.3 (3)	33	N.3 (3)
5	34	N.3 (4)	34	N.3 (3)
5	35	N.3 (4)	35	N.3 (3)
5	58	N.3 (3)	58	N.3 (4)
5	40	G.1 (2)		
5	49	M.1 (3)		
5	11	M.2 (2)		
5	23	M.2 (2)		
5	44	M.2 (2)		
5	51	M.2 (3)	51	M.2 (3)
5	39	D.1 (2)		
5	45	D.1 (2)		
5			47	D.1 (2)
6	41	N.2 (4)	41	N.2 (3)
6	29	N.3 (4)	29	N.3 (4)
6	30	N.3 (4)	30	N.3 (4)
6	32	N.3 (2)		
6	33	N.3 (2)		
6	34	N.3 (2)	34	N.3 (3)
6	36	N.3 (2)		
6	43	N.3 (2)		
6	46	N.3 (2)		
6	47	N.3 (2)		
6	54	N.3 (2)		

Table 2 (continued)
Items Coded to Generic Grade-level Expectations by More Than One Reviewer, Missouri Alignment Analysis for Mathematics, Grades 3–8 and 10, 2006 and 2007 Assessments

Grade	2006		2007	
	Assessment Item Number	Generic GLE (Number of Reviewers)	Assessment Item Number	Generic GLE (Number of Reviewers)
6	49	G.1 (2)		
6	57	M.2 (2)		
6	37	D.1 (2)	37	D.1 (3)
6	38	D.1 (2)	38	D.1 (3)
7			29	N.3 (3)
7	32	N.3 (2)	32	N.3 (2)
7			34	N.3 (3)
7			35	N.3 (3)
7			54	A.1 (2)
7			56	A.1 (2)
7			37	A.2 (3)
7			24	M.2 (2)
7	50	M.2 (2)	50	M.2 (3)
7			51	M.2 (3)
7			52	M.2 (2)
7	62	M.1 (2)	62	M.2 (2)
7			38	D.1 (3)
7			39	D.1 (3)
7			40	D.1 (2)
8	40	A.1 (2)	40	A.1 (3)
8			59	A.4 (2)
8			5	G.1 (2)
8			47	M.1 (2)
8			14	M.2 (3)
8	46	M.2 (2)	46	M.2 (3)
8			57	M.2 (2)
8	58	M.1 (2)	58	M.2 (3)
8			63	M.2 (2)
8			64	M.2 (2)
8			39	D.3 (2)
8	43	D.2 (2)		

Table 2 (continued)
Items Coded to Generic Grade-level Expectations by More Than One Reviewer, Missouri Alignment Analysis for Mathematics, Grades 3–8 and 10, 2006 and 2007 Assessments

Grade	2006		2007	
	Assessment Item Number	Generic GLE (Number of Reviewers)	Assessment Item Number	Generic GLE (Number of Reviewers)
10	56	N.1 (3)		
10	33	N.3 (2)	33	N.3 (3)
10	34	N.3 (2)	34	N.3 (3)
10	55	A.2 (2)	55	A.2 (2)
10	58	A.2 (3)	58	A.2 (2)
10	53	G.3 (2)		
10	12	M.2 (3)	12	G.1 (2)
10	17	M.2 (2)		
10	41	M.2 (2)		
10	48	M.2 (3)		
10	59	M.2 (3)		
10	40	D.2 (2)		
10	39	D.3 (3)		
10			4	G.1 (2)
10			8	G.1 (3)

Alignment of Curriculum Standards and Assessments

The 2006 and 2007 assessments for grades 3–8 and 10 were comprised of from 60 to 65 items (Table 3). Most items were 1-point multiple-choice items; from 7 to 11 items were constructed-response items worth 2 or 4 points each. The total point value ranged from 67 (grade 3) to 77 (grade 4) points. No field test items were included on the test or in the analysis.

The results of the analysis for the 2006 mathematics assessment for each of the four alignment criteria are summarized in Table 4. More detailed data on each of the criteria are given in the Appendix B in the first three tables. In Table 4, “YES” indicates that an acceptable level was attained between the assessment and the standard on the criterion. “WEAK” indicates that the criterion was nearly met, within a margin that could simply be due to error in the system. “NO” indicates that the criterion was not met by a noticeable margin—10% over an acceptable level for Depth-of-Knowledge Consistency, 10% over an acceptable level for Range-of-Knowledge Correspondence, and .1 under an index value of .7 for Balance of Representation.

Table 3
Number of Items and Point Value by Grade for Missouri Assessments, Grades 3–8 and 10, 2006 and 2007 Study

Grade Level	Number of Items	Number of Multi-Point Items	Constructed Response Items	Total Point Value
3	60	53	7	67
4	65	55	10*	77
5	62	54	8	70
6	61	54	7	68
7	62	55	7	69
8	64	54	10*	76
10	61	50	11*	74

* Includes one performance event assigned four points.

Alignment Results for the 2006 Assessments

Results from the analysis of the four alignment criteria indicate that the alignment was good for six of the seven grades for the 2006 mathematics assessment. There was full alignment for grades 4 and 6 as indicated by the fact that the standards and assessments met an acceptable level on each of the four alignment criteria for each of the five mathematics standards. The alignment only needs slight improvement for grades 3, 7, 8, and 10. In general, failing to meet an acceptable level on the Depth-of-Knowledge Consistency criterion for all five standards was the main shortcoming with respect to achieving full alignment between the assessment and the standards. The alignment at each grade is discussed in more detail.

Grade 3

The grade 3 assessment for 2006 and the standards were nearly fully aligned, except for the Balance-of-Representation criterion for Standard D (Data and Probability). Reviewers judged that nine items corresponded to three of the grade-level expectations under D, but seven of the items measured content related to one GLE (D.1.c.3 Read and interpret information from line plots and graphs), while one item corresponded to each of two other GLEs. This is not considered a major issue, since all of the other criteria were fully met. Balance for Data and Probability could be attained by replacing two of the items that correspond to the GLE D.1.c.3 by items that correspond to the other three GLEs under that standard.

Grade 4

The grade 4 2006 mathematics assessment and the standards were judged to be fully aligned.

Table 4
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grades 3-8 and 10, Standards and Assessments for Missouri Alignment Analysis

Table 4.1
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 3, Standards and Assessments for Missouri Alignment Analysis, 2006 Study

Grade 3	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Represent ation*
N - Numbers and Operations	YES	YES	YES	YES
A – Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M - Measurement	YES	YES	YES	YES
D - Data and Probability	YES	YES	YES	NO

Table 4.2
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 4, Standards and Assessments for Missouri Alignment Analysis, 2006 Study

Grade 4	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Represent ation*
N - Numbers and Operations	YES	YES	YES	YES
A - Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M - Measurement	YES	YES	YES	YES
D - Data and Probability	YES	YES	YES	YES

Grade 5

The alignment between the grade 5 2006 mathematics assessment and the standards needs slight improvement (Table 4.3). The main alignment issue is that too high a percentage of the items that target content under three of the standards has a depth-of-knowledge level that is below the level of the corresponding GLE. In addition, the analysis indicated that the Balance was weak for Standards N and G, and 20% of the items were coded to generic GLEs. The changes needed to achieve full alignment are relatively minor, except for fixing the large number of items coded to generic GLEs. Five items would need to be replaced to attain an acceptable level on depth of knowledge. Three items corresponding to Standard N, one item corresponding to Standard A, and one item corresponding to Standard D would need to be replaced by items with a DOK level

of 2 or higher to attain an acceptable level on the Depth-of-Knowledge Consistency criteria. If some of the items targeting Standard N were replaced by items that measure GLEs that are not overemphasized and with an appropriate DOK level, then this would improve the Balance. The main reason for the low Balance for Standard N is that too many of the items were coded to the generic GLE N3 because they corresponded to GLEs in lower grades. The weak Balance for Standard G would be improved by replacing two or three of the items corresponding to GLE G.2.a.5 with items that target other GLEs under that standard. Overall, full alignment at grade 5 could be attained by replacing seven or eight of the existing items.

Table 4.3
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 5, Standards and Assessments for Missouri Alignment Analysis, 2006 Study

Grade 5	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	NO	YES	WEAK
A - Algebraic Relationships	YES	WEAK	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	WEAK
M – Measurement	YES	YES	YES	YES
D - Data and Probability	YES	WEAK	YES	YES

Grade 6

The grade 6 2006 mathematics assessment and the standards were judged to be fully aligned.

Table 4.4
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 6, Standards and Assessments for Missouri Alignment Analysis, 2006 Study

Grade 6	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	YES	YES	YES
A - Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M – Measurement	YES	YES	YES	YES
D - Data and Probability	YES	YES	YES	YES

Grade 7

The grade 7 mathematics assessment for 2006 and five standards is judged to be reasonably aligned. The only alignment issue found was a weak Balance for Standard N. However, since all of the other alignment criteria were full met, weak Balance in one standard is not considered an issue. One reviewer, but not the same reviewer, coded a large number of the items to GLE N.1.b.7. However, the items coded by the majority of the reviewers were adequately spread over the GLEs under Standard N. At grade 7, the alignment is good.

Table 4.5
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 7, Standards and Assessments for Missouri Alignment Analysis, 2006 Study

Grade 7	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	YES	YES	WEAK
A - Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M - Measurement	YES	YES	YES	YES
D - Data and Probability	YES	YES	YES	YES

Grade 8

The alignment at grade 8 between the 2006 mathematics assessment and the five standards is reasonable. Only two alignment issues were found: Not a high enough proportion of the items targeting Standard D had a DOK level that was the same as or higher than the DOK level of the corresponding GLE. At grade 8, four of the six GLEs under Standard D were judged to have a DOK level of 3 (Sstrategic reasoning). However, nearly all of the items corresponding to GLEs under Standard D were judged to be at DOK Level 2, or between 2 and 3. Item 27 is an example of an item that was a good match. Reviewers judged that Item 27 had a DOK level between 2 and 3 and corresponded to GLE D.2.a.8 with a DOK level of 2. Even though four of the six GLEs under Standard D had a DOK level of 3, a majority of reviewers did not judge that any of the items corresponding to Standard D were at DOK Level 3. Under Standard N, GLE N.3.c.8 was overemphasized compared to the other GLEs. This is not considered as critical, since all of the other alignment criteria were fully met for Standard N. Full alignment could be attained by replacing at least 2 of the 15 items that correspond to Standard D with items at DOK Level 3 and replacing one or two items targeting N.3.c.8 with items that correspond to less emphasized GLEs under Standard N.

Table 4.6
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 8, Standards and Assessments for Missouri Alignment Analysis, 2006 Study

Grade 8	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	YES	YES	WEAK
A - Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M - Measurement	YES	YES	YES	YES
D - Data and Probability	YES	NO	YES	YES

Grade 10

The alignment between the grade 10 mathematics 2006 assessment and standards is reasonable. As was the case for grade 8, only one standard (Standard N) did not meet an acceptable level on the Depth-of-Knowledge Consistency criterion and one standard (Standard M) had a weak Balance. Full alignment could be attained by replacing two items with a DOK level of 1 that target Standard N with items at DOK Level 2. Items 33, 35, and 36 would be likely items to replace. Items 30, 41, 43, 45, and 51 were all judged to have a DOK level of 2 and are examples of the level of complexity that is needed. The weak Balance for Standard M, although not a major concern, could be removed by replacing one item with an item that corresponds to a GLE with fewer items, such as M.2.b.10 or M.2.d.10. The Balance would also be improved by replacing items assigned to the generic GLE M.2, such as Items 12, 17, or 41. Thus, full alignment could be attained at grade 10 by replacing three to four items.

Table 4.7
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 10, Standards and Assessments for Missouri Alignment Analysis, 2006 Study

Grade 10	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	NO	YES	YES
A - Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M - Measurement	YES	YES	YES	WEAK
D - Data and Probability	YES	YES	YES	YES

Alignment Results for the 2007 Assessments

The alignment of the 2007 Mathematics Assessments with the Missouri Mathematics Standards, similar to the alignment for 2006, is generally good (Table 5). For grades 3 and 4, the assessments and standards are fully aligned. For grades 5, 6, 7, and 8, the alignment is reasonable. For grade 10, the alignment needs slight improvement. As for 2006, the main alignment issue for 2007 is with Depth of Knowledge and Balance. The alignment for each grade level is discussed below.

Grade 3

The grade 3 2007 mathematics assessment and the standards were judged to be fully aligned.

Grade 4

The grade 4 2007 mathematics assessment and the standards were judged to be fully aligned.

Table 5
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grades 3–8 and 10, Standards and Assessments for Missouri Alignment Analysis

Table 5.1
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 3, Standards and Assessments for Missouri Alignment Analysis, 2007 Study

Grade 3	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	YES	YES	YES
A - Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M - Measurement	YES	YES	YES	YES
D - Data and Probability	YES	YES	YES	YES

Table 5.2
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 4, Standards and Assessments for Missouri Alignment Analysis, 2007 Study

Grade 4	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	YES	YES	YES
A - Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M - Measurement	YES	YES	YES	YES
D - Data and Probability	YES	YES	YES	YES

Grade 5

The alignment of the grade 5 mathematics assessment for 2007 and the Missouri standards is reasonable and much stronger than for 2006. Only one alignment issue was found for 2007, the Depth-of-Knowledge Consistency criterion was only weakly met for Standard D, Data and Probability. This alignment issue could be resolved by replacing one existing item, or adding one item with an assessment item at DOK Level 3. The majority of reviewers indicated that most of the 10 assessment items corresponding to Standard D had a DOK level of 2. However, half of the six GLEs under grade 5 Standard D have a DOK level of 3. Item 26 was rated with the highest DOK level (2.25 on the average).

Table 5.3
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 5, Standards and Assessments for Missouri Alignment Analysis, 2007 Study

Grade 5	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	YES	YES	YES
A - Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M – Measurement	YES	YES	YES	YES
D - Data and Probability	YES	WEAK	YES	YES

Grade 6

The alignment of the grade 6 2007 assessment for mathematics and the standards is reasonable. Full alignment could be attained by replacing one existing item, or adding one item that targets content related to Standard A with a DOK level of at least 2. Items 6

and 20 are possible items to consider for replacement because their DOK levels are lower than the DOK level of the corresponding objective. The Balance issue for Standard G could be resolved by replacing one item that currently targets GLE G.1.a.6 with one the targets one of the other GLEs under Standard G.

Table 5.4
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 6, Standards and Assessments for Missouri Alignment Analysis, 2007 Study

Grade 6	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	YES	YES	YES
A - Algebraic Relationships	YES	WEAK	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	WEAK
M - Measurement	YES	YES	YES	YES
D - Data and Probability	YES	YES	YES	YES

Grade 7

At grade 7, the alignment between the 2007 mathematics assessment and the standards is reasonable. The alignment issues for Standard N could be resolved by replacing one item, or adding one item with a higher DOK level. However, all of the items that relate to Standard N come from Session 2 and are used for the norm-referenced score. Whereas most of the GLEs under Standard N at grade 7 were judged to have a DOK level of 2, most of the corresponding items were judged to be at DOK Level 1. The weak alignment for Standard N is not considered a major concern because it could be fixed by replacing or adding one item. The weak Depth-of-Knowledge issue for Standard D could be improved by replacing or adding two items. The new items should have a DOK level of at least 2 or higher, depending on the complexity of the targeted GLE. Overall, full alignment could be attained at grade 7 by replacing or adding a total of three items.

Table 5.5
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 7, Standards and Assessments for Missouri Alignment Analysis, 2007 Study

Grade 7	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	WEAK	WEAK	WEAK
A - Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M - Measurement	YES	YES	YES	YES
D - Data and Probability	YES	WEAK	YES	YES

Grade 8

The alignment of the grade 8 2007 assessment for mathematics and the standards is reasonable. Full alignment could be attained by replacing one existing item, or adding one item that targets content related to Standard D with a DOK level of at least 2. The Balance issue for Standard A could be resolved by replacing one item that currently targets GLE A.2.a.8 with one that targets one of the other GLEs under Standard A. Thus, only two items need to be replaced or added to attain full alignment, using the acceptable levels as defined for this study.

Table 5.6
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 8, Standards and Assessments for Missouri Alignment Analysis, 2007 Study

Grade 8	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	YES	YES	YES
A - Algebraic Relationships	YES	YES	YES	WEAK
G - Geometric and Spatial Relationships	YES	YES	YES	YES
M - Measurement	YES	YES	YES	YES
D - Data and Probability	YES	WEAK	YES	YES

Grade 10

The alignment of the grade 10 2007 assessment for mathematics and the standards needs slight improvement. At least two items need to be replaced or added by items with a higher DOK level for both Standards N and G. As for the other grades, most of the items corresponding to Standard N are in Session 2 of the test, the norm-referenced part

of the assessment. The Number and Operations items in Session 2 of the assessment tend to a have lower DOK level than required by the Missouri Standards. It may be necessary to add two items in order to have an appropriate distribution of items by DOK level. For Standard G, Item 3 could be replaced by an item with a DOK level of 3 and Item 28 could be replaced by an item at DOK Level 2 in order to meet an acceptable level on the Depth-of-Knowledge Consistency criterion. The Balance issue for the Standards N and M could be removed by replacing items that currently correspond to N.3.e.10 and M.2.c.10 with items that measure content related to less emphasized GLEs under each standard. Overall, a total of five or six items need to be replaced or added to attain full alignment.

Table 5.7
Summary of Acceptable Levels on Alignment Criteria for Mathematics Grade 10, Standards and Assessments for Missouri Alignment Analysis, 2007 Study

Grade 10	***Alignment Criteria***			
Standards	*Categorical Concurrence*	*Depth-of-Knowledge Consistency*	*Range of Knowledge*	*Balance of Representation*
N - Numbers and Operations	YES	NO	YES	WEAK
A - Algebraic Relationships	YES	YES	YES	YES
G - Geometric and Spatial Relationships	YES	NO	YES	YES
M – Measurement	YES	YES	YES	NO
D - Data and Probability	YES	YES	YES	YES

Reviewers' Comments

Reviewers were instructed to document any Source-of-Challenge issue and to provide any other comments they may have. These comments can be found in Tables (grade).5 and (grade).7 in Appendix C (2006) and Appendix F (2007). After coding each grade-level assessment, reviewers also were asked to respond to five debriefing questions. All of the comments made by the reviewers are given in Appendix D (2006) and Appendix G (2007). The notes in general offer an opinion on the item or give an explanation of the reviewers' coding.

Reliability Among Reviewers

The overall intraclass correlation among the mathematics reviewers' assignment of DOK levels to items was reasonable for four reviewers for both the 2006 and 2007 assessments (Table 6). An intraclass correlation value greater than 0.8 generally indicates a high level of agreement among the reviewers. All 14 intraclass correlations in assigning a DOK level to items were above .7. This indicates reasonable agreement. A pairwise comparison is used to determine the degree of reliability of reviewer coding at the grade-level expectation level and at the standard level. The standard pairwise comparison values are high, all above .8, except for the grade 10 2007 assessment. The GLE pairwise agreement was a little lower than for most alignment studies, below .6. Reviewers'

agreement in assigning items to GLEs for grade 10 (2007) was very low, .30. The low reviewer agreement in assigning items to GLEs could have an impact on the Balance of Representation results. Thus, the results for grade 10 Balance need to be reviewed carefully and changes made only if warranted. The lower agreement among reviewers in assigning items to GLEs can be an indication of the lack of clear GLE statements, or some ove lap among the GLEs, as much as the need for more reviewer training. The relatively high values for the intraclass correlation and in assigning items to standards suggest that the training was good.

Table 6
Intraclass and Pairwise Comparisons, Missouri Alignment Analysis for Mathematics Grades 3–8 and 11, Assessments, 2006 and 2007 Study

Grade	Intraclass Correlation	Pairwise Comparison:	Pairwise: Grade level expectation	Pairwise: Standard
2006				
3	.83	.71	.57	.83
4	.85	.70	.44	.82
5	.75	.69	.56	.86
6	.76	.67	.56	.84
7	.72	.67	.54	.85
8	.75	.51	.45	.83
10	.73	.69	.41	.81
2007				
3	.75	.67	.55	.84
4	.78	.69	.56	.85
5	.75	.65	.47	.86
6	.72	.69	.74	.88
7	.75	.68	.52	.90
8	.86	.74	.50	.88
10	.73	.54	.30	.77

Summary

Eight reviewers analyzed the mathematics assessments for 2006 and 2007 and the Missouri Standards at a three-day institute held September 27-29, 2009, in Columbia, Missouri. The reviewers included mathematics content experts, district mathematics supervisors, mathematics teachers, and mathematics education doctoral students. Half of the reviewers were from Missouri and half of them were from other states. All eight reviewers analyzed the depth-of-knowledge level of the grade-level expectations under each of the five standards for all grades 3–8 and 10. At least four reviewers analyzed the assessment for each year. The intraclass correlation among the reviewers across all 14 analyses was reasonably high, along with the pairwise agreement in assigning items to standards. However, pairwise agreement in the assignment of items to GLEs was lower

than for other alignment studies. This lower level of agreement could be the result of lack of clarity in the GLE statements, or the need for more training of the reviewers.

For both years, 2006 and 2007, the alignment was reasonable. The alignment needed slight improvement in only one grade for each year—grade 5 for 2006 and grade 10 for 2007. From five to seven items would need to be replaced or added to achieve full alignment on these two test forms. For all other grades for each of the years, the assessment and the standards were either fully aligned or reasonably aligned. For these assessment forms and grades, less than five items would need to be replaced or added. The assessment and standards were found to be fully aligned for grades 4 and 6 (2006) and for grades 3 and 4 (2007). All test forms for all grades and years had a sufficient number of items for each standard that were, in general, adequately distributed among the GLEs. The main alignment issues were that one or two of the items had too low a depth-of-knowledge level compared to the DOK level of the corresponding GLE. Also, for a few standards, one GLE was overemphasized compared to the other GLEs under that standard. A relatively higher number of items were coded to generic GLEs, indicating a lack of precise match of the item to a specific GLE. Most of these items were on the norm-referenced part of the assessment, Session 2, which would be more difficult to change. Considering all seven grades and both years, the alignment is judged to be reasonable.

References

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison: University of Missouri, Missouri Center for Education Research.